Tensor Query Expansion: A cognitively motivated relevance model

Mike Symonds, Peter D. Bruza, Laurianne Sitbon, Ian Turner
Faculty of Science and Technology,
Queensland University of Technology, Brisbane, Qld, Australia
m.symonds@student.qut.edu.au, (p.bruza, l.sitbon, i.turner)@qut.edu.au

Abstract *In information retrieval, a user's query is often not a complete representation of their real information need. The user's information need is a cognitive construction, however the use of cognitive models to perform query expansion have had little study.*

In this paper, we present a cognitively motivated query expansion technique that uses semantic features for use in ad hoc retrieval. This model is evaluated against a state-of-the-art query expansion technique. The results show our approach provides significant improvements in retrieval effectiveness for the TREC data sets tested.

Keywords Information storage and retrieval, query expansion, natural language processing, tensors

1 Introduction

Information retrieval researchers have known that a user's query is typically an imprecise description of the user's real information need ever since the Cranfield experiments in document retrieval in the 1960's. This is all the more relevant today with web queries, which are commonly between two and three words in length. For these reasons there has been, and still is, a strong interest in the use of query expansion techniques. These techniques augment the original query, creating what is hoped to be a more accurate representation of the user's real information need and has been shown to consistently increase retrieval effectiveness [14].

Current state-of-the-art query expansion techniques are often based on word statistics found within documents and ignore information about term dependencies that are inherent to natural language [7, 20]. However, there is growing evidence that suggests query expansion techniques that use term dependency information, such as word proximity and word co-occurrences, can provide more effective query expansion over these traditional approaches [9, 10]. Very few of these dependency based techniques are motivated from a cognitive perspective, from which the user's real information need is created [11].

In this paper we present a formal query expansion approach, we call *tensor query expansion* (TQE), based on a cognitively motivated model of word meaning. We hypothesise that as this approach is more intuitively linked to the cognitive construct of a user's real information need, a word meaning model could be used within the query expansion process to make query representations more like the user's real information need. Our results demonstrate significant improvements in retrieval effectiveness, and support the argument that the user's real information need can be successfully modelled using a cognitive model of word meaning.

Proceedings of the 16th Australasian Document Computing Symposium, Canberra, Australia, 2 December 2011.

2 Related Work

The main areas of research that provide a theoretical framework for our approach include: (i) the use of query expansion techniques to improve the representation of a user's information need, (ii) linguistic theories of word meaning, and (iii) the use of semantic spaces to model word meaning.

2.1 Query Expansion Techniques

State-of-the-art document retrieval models, such as those from the language modelling group, are framed within probabilistic settings, with documents and queries represented as statistical distributions.

The language modelling framework does not have a natural extension for query expansion. Ad hoc approaches have been applied with some success [1, 3]. However, more formal techniques, including Zhai and Lafferty's model-based feedback and Lavrenko and Croft's relevance models are often used [20, 7]. The unigram relevance model is often used as a state-of-the-art benchmark for research into query expansion techniques [10].

The relevance modelling approach involves estimating the probability of observing a word w given some relevant evidence for a particular information need, represented as query Q. The relevance model is a multinomial distribution in which the conditional probability is computed as:

$$P(w|Q) = \int_D P(w|D)P(D|Q). \qquad (1)$$

By assuming that most of the relevant information comes from the set of relevant documents for a query

Q, the conditional probability estimate of equation (1) can be rewritten as:

$$P(w|R) \approx \frac{\sum_{D \in \mathcal{R}_Q} P(w|D)P(Q|D)P(D)}{\sum_w \sum_{D \in \mathcal{R}_Q} P(w|D)P(Q|D)P(D)}, \quad (2)$$

where $\mathcal{R}_Q$ is the set of documents (pseudo) relevant to query Q. To simplify the estimation, there is an assumption that $P(D)$ is uniform over this set of documents. By using a set of the most probable terms from this distribution the query representation is updated, often through linear interpolation with the original query model, and used in re-ranking the documents, as shown:

$$P(w|Q) = \lambda P_o(w|Q) + (1 - \lambda)P(w|R), \quad (3)$$

where λ is the feedback interpolation coefficient that determines the mix with the original estimate $P_o(w|Q)$. In the unigram case, the relevance model estimates are often based on the Dirichlet smoothed term likelihood scores, expressed as:

$$P(w|D) = \frac{df_w + \mu \frac{cf_w}{|C|}}{|D| + \mu}, \quad (4)$$

where df_w is the document frequency of term w, cf_w is the collection frequency of term w, $|C|$ is the word count in the collection, $|D|$ is the word count of the document and μ is the Dirichlet smoothing parameter. Within this paper, this instance of the relevance model is referred to as RM3.

Our approach works within the relevance modelling framework and replaces $P(w|R)$ in equation (3) with an analogous estimate produced by a formal model of word meaning. The retrieval effectiveness of our cognitively motivated relevance model is compared with the unigram based relevance model on a number of newswire data sets. This comparison is centered around the use of word meanings within our cognitive approach.

2.2 Word Meaning

The structuralist theories of linguistics, championed by Ferdinand de Saussure (1916), stated that meaning arose from the relationships between words and provided a relatively clean linguistic framework, free of psychology, sociology and anthropology [4]. Based on these ideas, word meaning was created by two types of relationships: (i) syntagmatic and (ii) paradigmatic associations.

A syntagmatic association exists between two words if they co-occur more frequently than expected from chance. Some common examples may include "coffee-drink" and "sun-hot" [13]. These associations can also have varying strengths. Consider the example sentence "A dog bit the mailman". The term "dog" would likely have a stronger syntagmatic association with "bit" than "mailman", based on the fact that the word "bit" would likely co-occur with "dog" more often.

A paradigmatic association exists between two words if they can substitute for one another in a sentence without affecting the grammaticality or acceptability of the sentence. Some common examples may include "drink-eat" and "quick-fast" [13]. In the example sentence, "A dog bit the mailman", the word "bit" could be replaced with "chased", hence "bit" and "chased" could be said to have a paradigmatic association.

2.3 Semantic Space Models

Linked to structuralist ideas of linguistics, researchers have argued that relationships between words can be modelled by comparing the distributions of words within text [16]. A popular approach to representing these word distributions is to collect word occurrence frequencies and place them in high-dimensional *context* vectors [18]. This approach allows techniques from linear algebra to be used to model relationships between objects, including semantic associations when a word space is developed.

There have been a number of successful psychologically relevant semantic space models that learn semantic associations directly from text, including HAL (Hyperspace Analogue to Language [8]) and LSA (Latent Semantic Analysis [6]). More recent models demonstrate that encoding structural information into the semantic space improves performance on a number of cognitive tasks [5, 15, 17] and can help address weaknesses raised by the lack of structural information in models like LSA [12].

Of these more recent models, the tensor encoding (TE) model [17] provides measures of syntagmatic and paradigmatic associations between words. We argue that this strong connection to structural linguistic theory creates a solid foundation for modelling the creation of word meanings, that likely form part of the user's cognitive process when developing their real information need, for use in an information retrieval task.

3 The Tensor Encoding model

The TE model creates a semantic space based on tensor representations of words. These tensor representations are built by encoding the word order and word co-occurrence information found in natural language. The tensor order created within the space depends on the size of the tuples encoded in the vocabulary binding process.

To gain a better understanding of how the TE model creates tensors that formally capture this word-order and co-occurrence information, consider the second order TE model created for the example sentence, "*A dog bit the mailman*", where *A* and *the* are considered to be stop words (noisy, low information terms that are not included in the vocabulary). Each term in the vocabulary is first assigned an environment vector, which corresponds to the unit vector for the term's id value:

Term-id	Term	Environment vector
1	dog	$\boldsymbol{e}_{dog} = (1 \quad 0 \quad 0)^T$
2	bit	$\boldsymbol{e}_{bit} = (0 \quad 1 \quad 0)^T$
3	mailman	$\boldsymbol{e}_{mailman} = (0 \quad 0 \quad 1)^T$

A *memory tensor* for each term is built by summing the proximity-scaled Kronecker products of the environment vectors within a sliding context window over the text. For the second order TE model, memory matrices are created by the binding process defined by:

$$\boldsymbol{M}_w = \sum_{k \in CW}^{k \prec w} (R - d_k + 1).\boldsymbol{e}_k \otimes \boldsymbol{e}_w^T + \sum_{k \in CW}^{k \succ w} (R - d_k + 1).\boldsymbol{e}_w \otimes \boldsymbol{e}_k^T, \quad (5)$$

where w is the target term, k is a non-stop word found within the sliding context window (CW), $k \prec w$ indicates that term k appears before term w in the context window, $k \succ w$ indicates that term k appears after term w, R is the radius of the sliding context window, and d_k is the distance between term k and term w. Note, stop words are not bound, but they are included when determining the window boundaries. Consider the memory matrices created for the vocabulary terms using a sliding context window with radius 2.

Binding Step 1: $\overbrace{A_s \quad [dog] \quad bit \quad the_s} \quad mailman$

$$\begin{aligned}\boldsymbol{M}_{dog} &= 2 \times \boldsymbol{e}_{dog} \otimes \boldsymbol{e}_{bit}^T \\ &= 2 \times \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (0 \quad 1 \quad 0) = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.\end{aligned}$$

Binding Step 2: $\overbrace{A_s \quad dog \quad [bit] \quad the_s \quad mailman}$

$$\begin{aligned}\boldsymbol{M}_{bit} &= 2 \times \boldsymbol{e}_{dog} \otimes \boldsymbol{e}_{bit}^T + \boldsymbol{e}_{bit} \otimes \boldsymbol{e}_{mailman}^T \\ &= 2 \times \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} (0\,1\,0) + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} (0\,0\,1) \\ &= \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.\end{aligned}$$

Binding Step 3: $A_s \quad dog \quad \overbrace{bit \quad the_s \quad [mailman]}$

$$\begin{aligned}\boldsymbol{M}_{mailman} &= \boldsymbol{e}_{bit} \otimes \boldsymbol{e}_{mailman}^T \\ &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} (0 \quad 0 \quad 1) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.\end{aligned}$$

The resulting pattern is that all non-zero elements are situated on the row or column corresponding to the target term's term-id. If this vocabulary building process was performed over the entire corpus the general form of a *memory matrix* would be similar to:

$$\boldsymbol{M}_w = \begin{pmatrix} 0, \quad \ldots, 0, \quad f_{1w}, \quad 0, \quad \ldots, 0 \\ \cdots \\ 0, \quad \ldots, 0, \quad f_{(w-1)w}, \quad 0, \quad \ldots, 0 \\ f_{w1}, \ldots, f_{w(w-1)}, f_{ww}, f_{w(w+1)}, \ldots, f_{wN} \\ 0, \quad \ldots, 0, \quad f_{(w+1)w}, \quad 0, \quad \ldots, 0 \\ \cdots \\ 0, \quad \ldots, 0, \quad f_{Nw}, \quad 0, \quad \ldots, 0 \end{pmatrix},$$

where f_{iw} is the value in row i column w of the matrix which represents the proximity scaled co-occurrence frequencies of term i before term w, f_{wj} is the value in row w column j of the matrix that represents the proximity scaled co-occurrence of term j after term w, and N is the number of unique terms in the vocabulary.

It is worth noting that the illustrative example provided was for the second order implementation of the TE model, and hence the representations are matrices. The TE model can be extended to form higher order tensors that hold n-tuple information by modifying equation (5) to sum the Kronecker products of n-tuples. For this research we chose to use the second order TE model. Using higher order TE models is left for future work.

The sparse memory matrices of the second order TE model provide opportunities for efficient construction, as outlined in [17]. They also allow for efficient evaluation measures to be developed to achieve the goal of modelling word meaning.

3.1 Computing Word Meaning

Based on the structuralist theories of linguistics the TE model aims to extract the strength of syntagmatic and paradigmatic associations between terms to allow it to construct word meaning.

3.1.1 Syntagmatic associations:

Due to the unique structure of the memory matrices, it was shown in [17] that the strength of syntagmatic associations between a sequence of priming terms $Q = (q_1, \ldots, q_p)$ and any vocabulary term w, can be efficiently calculated by measuring the cosine of the angle θ between the memory matrices for Q ($\boldsymbol{M_Q}$) and w ($\boldsymbol{M_w}$):

$$s_{\text{syn}}(Q, w) = \cos\theta = \frac{< \boldsymbol{M_Q}, \boldsymbol{M_w} >}{||\boldsymbol{M_Q}||_F ||\boldsymbol{M_w}||_F}, \quad (6)$$

where

$$< \boldsymbol{M_Q}, \boldsymbol{M_w} > = \sum_{\substack{j=1 \\ w \in Q}}^{N} s_w^2 f_{jw}^2 + \sum_{\substack{j=1 \\ j \neq w \\ w \in Q}}^{N} s_w^2 f_{wj}^2 + \sum_{\substack{i=q_1 \\ i \neq w}}^{q_m} (s_i^2 f_{wi}^2 + s_i^2 f_{iw}^2),$$

and

$$||\boldsymbol{M_Q}||_F = \sqrt{\sum_{i=q_1}^{q_m} \left[\sum_{j=1}^{N} s_i^2 f_{ji}^2 + \sum_{\substack{j=1 \\ j \neq i}}^{N} s_i^2 f_{ij}^2 \right]},$$

and

$$||\boldsymbol{M_w}||_F = \sqrt{\sum_{j=1}^{N} f_{jw}^2 + \sum_{\substack{j=1 \\ j \neq w}}^{N} f_{wj}^2},$$

and where the memory matrix for Q was constructed by summing the memory matrices of the individual terms

in the sequence $\boldsymbol{M_Q} = \boldsymbol{M_{q_1}} + \ldots + \boldsymbol{M_{q_p}}$, $q_1, \ldots, q_m$ are the list of m unique priming terms found in Q having $m \leq p$, s_i is the number of times term q_i appears in Q, f_{ab} is the co-occurrence frequency of term a appearing before term b in the vocabulary, f_{ba} is the co-occurrence frequency of term a appearing after term b.

This measure of syntagmatic association was shown to have linear time complexity and be effective at predicting words that are most likely to precede or succeed another word in text [17].

3.1.2 Paradigmatic associations:

Having explicit 2-tuple co-occurrence information stored in the memory matrices means that the TE model allows probabilistic measures to be used in addition to geometric measures. This means that information theoretic measures, like mutual information, could be used if they provided the best performance for a given task. However, for measuring the strength of paradigmatic associations between a sequence of priming terms $Q = (q_1, \ldots, q_p)$ and a vocabulary term w, the following measure was shown in [17] to perform better on a synonym judgement task than other recent semantic space models encoding structural information:

$$s_{\text{par}}(Q, w) = \frac{1}{Z_{\text{par}}} \sum_{j=q_1}^{q_p} \sum_{i=1}^{N} \frac{f_{ij} f_{iw} + f_{ji} f_{wi}}{f_j f_w}, \quad (7)$$

where f_j is the vocabulary frequency of term j, f_{ji} is the ordered co-occurrence frequency of term j before term i, N is the size of the vocabulary, and Z_{par} normalizes the scores, such that $\sum_{w \in V} s_{\text{par}}(Q, w) = 1$.

As we plan to combine the syntagmatic and paradigmatic measures for the task of query expansion, we chose to modify equation (8) to extract more pure paradigmatic associations, and hence reduce the scores of terms that co-occur with query terms (indicating syntagmatic relations). The resulting paradigmatic feature function was defined as:

$$s_{\text{par1}}(Q, w) = \frac{1}{Z_{\text{par1}}} \sum_{j=q_1}^{q_p} \sum_{i=1}^{N} \frac{f_{\overline{ij}} . f_{\overline{iw}}}{\max(f_{\overline{ij}}, f_{\overline{iw}}, f_{\overline{wj}})^2}, \quad (8)$$

where $f_{\overline{ij}} = (f_{ji} + f_{ij})$, $f_{\overline{iw}} = (f_{wi} + f_{iw})$, $f_{\overline{wj}} = (f_{jw} + f_{wj})$, N is the size of the vocabulary, max() returns the maximum argument value, and Z_{par1} normalizes the scores, such that $\sum_{w \in V} s_{\text{par1}}(Q, w) = 1$.

4 Tensor Query Expansion

Research into models of memory have demonstrated that the type of semantic information that is most useful to perform a given task varies [5, 17]. For example, on a synonym judgement task, paradigmatic associations are most helpful, while on a task estimating the most common pre-ceding or post-ceding term for a target term, the syntagmatic associations are most useful.

For the task of query expansion, we assume that the underlying word meanings that form the user's information need are likely formed by a mix of both syntagmatic and paradigmatic associations. This assumption will be tested by comparing the retrieval effectiveness achieved when different mixes of syntagmatic and paradigmatic associations are used to form estimates for query models within the relevance modelling framework on an ad hoc retrieval task.

As outlined in section 2.1, relevance models provide a formal method for query expansion within the language modelling framework. Equation (2) shows the relevance model process includes estimating the probability $P(w|R)$, of observing a word w based on relevant evidence, often (pseudo) relevant documents, for a particular query Q using a multinomial distribution. Our aim will be to create an analogous distribution to estimate $P(w|R)$. These estimates will be based on word meanings, more specifically the combination of syntagmatic and paradigmatic associations found with the vocabulary created by the TE model on a set of (pseudo) relevant documents. We call this query expansion technique, *tensor query expansion* (TQE).

To formally estimate the conditional probability we use a Markov random field, similar to that used in [10]. Let an undirected graph G contain nodes that represent random variables, and the edges define the independence semantics between the random variables. Within the graph, a random variable is independent of its non-neighbours given observed values of its neighbours.

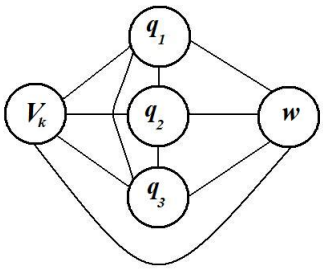

Figure 1: Example of the TQE graphical model for a three term query.

Figure 1 shows a graph G that consists of query nodes q_i, expansion term node w, and a vocabulary node V_k. Term w is constrained to exist within the vocabulary V_k, which is built from a set of k documents considered (pseudo) relevant to Q. We parameterize the graph based on clique sets to provide more flexibility in encoding useful features over cliques in the graph. The joint distribution over the random variables in G is defined by:

$$P_{G,\Gamma}(Q, w, V_k) = \frac{1}{Z_\Gamma} \prod_{c \in cl(G)} \varphi(c; \Gamma), \quad (9)$$

where $Q = q_1, \ldots, q_p$, $cl(G)$ is the set of cliques in G, each $\varphi(.; \Gamma)$ is a non-negative *potential function* over clique configurations parameterized by Γ,

and $Z_\Gamma = \sum_{Q,w} \prod_{c \in cl(G)} \varphi(c;\Gamma)$ normalizes the distribution. The joint distribution is uniquely defined by the graph G, potential functions φ and the parameter Γ. Using the fact that the logarithm of products is equal to the sum of logarithms, the simplified form of the joint distribution becomes:

$$\log P_{G,\Gamma}(Q,w,V_k) = \frac{1}{Z_\Gamma} \sum_{c \in cl(G)} \log \varphi(c;\Gamma), \quad (10)$$

where the potential functions are commonly parameterized as:

$$\varphi(c;\Gamma) = \exp[\gamma_c f(c)], \quad (11)$$

with $f(c)$ being some real-valued *feature function* over clique values and γ_c is the weight given to that particular feature function. Substituting equation (11) into equation (10) gives:

$$\log P_{G,\Gamma}(Q,w,V_k) = \frac{1}{Z_\Gamma} \sum_{c \in cl(G)} \gamma_c f(c). \quad (12)$$

After G is constructed, we can compute the conditional probability of an expansion term w given Q, as:

$$P_{G,\Gamma}(w|Q) = \frac{P_{G,\Gamma}(Q,w,V_k)}{\sum_{w \in V_k} P_{G,\Gamma}(Q,w,V_k)}, \quad (13)$$

where V_k is the universe of all possible vocabulary terms and w is a possible expansion term.

By using equation (12) and equation (13) with constant terms removed, a rank equivalent form for the conditional probability can be written as:

$$P_{G,\Gamma}(w|Q) \propto \sum_{c \in cl(G)} \gamma_c f(c), \quad (14)$$

where a constraint of $\sum_{c \in cl(G)} \gamma_c = 1$ is applied for ease of training.

4.1 Model Parameterization

The conditional probability expressed in equation (14), provides a formal method for combining feature functions, designed to extract various types of vocabulary term dependencies, mapped via cliques in the graph. For the graph shown in figure 1, a number of useful clique sets capturing dependencies are summarised in table 1.

Since it is not our goal to find optimal feature functions, but to demonstrate the use of a Markov random field to formally combine feature functions that model syntagmatic and paradigmatic associations, we focus on evaluating estimates over the clique sets relevant to the syntagmatic and paradigmatic measures.

To enable a more balanced comparison of the influence of each feature we first convert the syntagmatic measure to a distribution by normalising the scores, such that the feature function in equation (6) becomes:

Set	Description
T_{par}	Set of cliques containing the vocabulary node and exactly one query term node and the expansion term (w) node.
T_{syn}	Set of cliques containing the vocabulary node and exactly one query term node and the expansion term (w) node, with query term node and expansion term node connected by an edge.

Table 1: Summary of TQE clique sets to be used.

$$s_{\text{syn1}}(Q,w) = \frac{1}{Z_{\text{syn1}}} s_{\text{syn}}(Q,w), \quad (15)$$

where $Z_{\text{syn1}} = \sum_{w \in V_k} s_{\text{syn}}(Q,w)$ normalises the scores. Using the T_{syn} and T_{par} clique sets, and our feature functions $s_{\text{syn1}}(Q,w)$ and $s_{\text{par1}}(Q,w)$, equation (14) becomes:

$$P_{G,\Gamma}(w|Q) \propto \gamma_{T_{\text{syn}}} s_{\text{syn1}}(Q,w) + \gamma_{T_{\text{par}}} s_{\text{par1}}(Q,w), \quad (16)$$

where $\gamma_{T_{\text{syn}}}, \gamma_{T_{\text{par}}} \in [0,1]$ and $\gamma_{T_{\text{syn}}} + \gamma_{T_{\text{par}}} = 1$. By normalising the distribution and replacing $\gamma_{T_{\text{syn}}}$ and $\gamma_{T_{\text{par}}}$ with a single interpolation parameter, γ, the rank equivalent estimate in equation (16) can be rewritten as:

$$P_{G,\Gamma}(w|Q) = \frac{1}{Z_\Gamma} \left[\gamma s_{\text{syn1}}(Q,w) + (1-\gamma) s_{\text{par1}}(Q,w)\right], \quad (17)$$

where $\gamma \in [0,1]$, mixes the amount of syntagmatic and paradigmatic features used in the estimation, and $Z_\Gamma = \sum_{w \in V_k} [\gamma s_{\text{syn1}}(Q,w) + (1-\gamma) s_{\text{par1}}(Q,w)]$, is used to normalise the distribution.

As the estimate in equation (17) is considered analogous to the estimate $P(w|R)$ used in the relevance modelling framework, we argue that our $P_{G,\Gamma}(w|Q)$ can replace the $P(w|R)$ in equation (3) giving us a cognitively motivated method of updating the query model within the language modelling framework. It is worth noting that one of the major differences between the unigram based relevance model and our approach is that our estimates are based on the vocabulary measures, not the document statistics. Using the relevance models feedback interpolated form shown in equation (3), the final conditional probability becomes:

$$P(w|Q) = \lambda P_o(w|Q) + (1-\lambda) P_{G,\Gamma}(w|Q). \quad (18)$$

Our cognitively motivated relevance model needs to also be considered in terms of the computational costs of extracting these semantic features from the (pseudo) relevant document set.

5 Computational Complexity

The TQE technique uses two semantic features that measure the strength of syntagmatic and paradigmatic associations. The creation of the memory matrices in equation (5) provides a formalism for capturing the

co-occurrences and encoding word order. However, the original TE model research [17] demonstrated that the word order and co-occurrence information is efficiently captured within low dimension storage vectors (SV) due to the unique structure of the memory matrices.

The dimensionality of the storage vectors required is based on the size of the vocabulary created and the radius of the context window used in the vocabulary binding process.

For example, on a synonym judgement task using a vocabulary of 134,000 terms, the TE model's best performance was achieved using the paradigmatic measure, a context window of radius one and storage vectors of 1,000 dimensions [17]. This supports previous research [15] that showed paradigmatic associations are most effectively modelled when a very small context window is used. When this is considered alongside the fact that the vocabulary size created from the set of top 30 (pseudo) relevant documents in our ad hoc retrieval experiments was less than 10,000, a storage vector of 50 dimensions is chosen to model the paradigmatic associations within our TQE approach.

Considering the worst case time complexity of the paradigmatic feature in equation (7) is $T(n) = O(\frac{D^2_{SV_{\text{par}}}}{4}.|Q|)$, where $D_{SV_{\text{par}}}$ is the dimensionality of the storage vector (set to 50), and $|Q|$ is the length of the query, keeping the dimensionality of the storage vector small is important.

However, for tasks relying on syntagmatic associations, past research [19] has shown that using larger context windows leads to better performance. For this reason we chose to create a separate semantic space to model syntagmatic associations. This semantic space was built using a context window of radius 150 and storage vectors with 500 dimensions. Based on these dimensions the memory footprint of the storage vectors to build the two semantic spaces used by the TQE technique would be at most 21 MBytes ($550 \times 10,000$ integers), assuming a four byte integer.

The original TE model research [17] showed that the worst case time complexity of the syntagmatic feature, in equation (6) was $T(n) = O(\frac{D_{SV_{\text{syn}}}}{2}.|Q|)$, where $D_{SV_{\text{syn}}}$ is the dimensionality of the syntagmatic storage vector (set to 500 in the TQE approach).

6 Experimental Results

Evaluation of the TQE approach was performed on the TREC data sets outlined in table 2. The AP and WSJ data sets were chosen as they were likely to contain very different content from each other, and hence should form different strength semantic associations for the same queries.

A common approach for evaluating query expansion approaches is through the measure of average retrieval effectiveness and robustness on ad hoc retrieval tasks using pseudo-relevance feedback [10].

Name	Description	# Docs	Topics
AP	Assoc. Press 88-90	242,918	train: 1-150 test: 151-200
WSJ	Wall Street Journal 87-92	173,252	train: 1-150 test: 151-200

Table 2: Overview of TREC collections and topics

The experiments in this research were carried out using a modified version of the Lemur Toolkit[1]. All collections were stopped with the default 418 word Indri stop list and stemmed using a Porter stemmer. In all experiments, only the title component of the topics were used to construct the initial queries.

6.1 Ad Hoc Retrieval

The TQE approach was compared to a baseline unigram language model (noFB) and a unigram relevance model (RM3). The following ad hoc retrieval experiments use a manual train/test split, as outlined in the *Topics* column of table 2.

For RM3 the Dirichlet smoothing parameter, μ was trained and for the TQE approach, the mixing parameter γ was trained. Both TQE and RM3 were evaluated using 30 feedback documents and 30 expansion terms on all data sets. The mean average precision (MAP) for the top ranked 1000 documents are reported in table 3.

Test Set	noFB	RM3	TQE
AP (151-200)	0.2112	0.2495^{α}	$0.2683^{\alpha\beta}$
WSJ (151-200)	0.3244	0.3546^{α}	$0.3831^{\alpha\beta}$

Table 3: Mean average precision (MAP) scores for the unigram language model (noFB), unigram relevance model (RM3) and cognitively motivated relevance model (TQE). The superscripts α and β indicate statistically significant improvements using a two-tailed paired t-test ($p < 0.05$) over noFB and RM3 respectively.

The ad hoc retrieval results suggest that the TQE approach can significantly improve the average precision results when compared to both RM3 and the baseline language model (noFB) on the WSJ and AP data sets. To gain better insight into how the TQE retrieval effectiveness for each query compares to that of the unigram relevance model a robustness analysis is worthwhile.

6.2 Robustness

Robustness includes considering the ranges of relative increase/decrease in average precision and the number of queries that were improved/hurt, with respect to RM3. Figure 2 illustrates the robustness of the TQE average precision scores reported in table 3 on the AP and WSJ data sets. In each graph in figure 2 the test

[1]The Lemur toolkit for language modelling and information retrieval: http://www.lemurproject.org

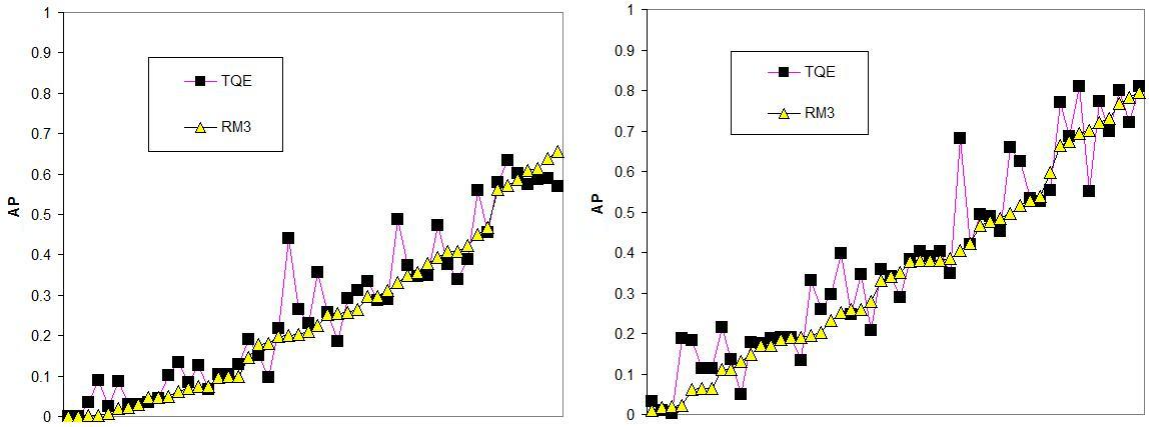

Figure 2: Robustness comparison of RM3 and TQE on the AP (left) and WSJ (right) test topics.

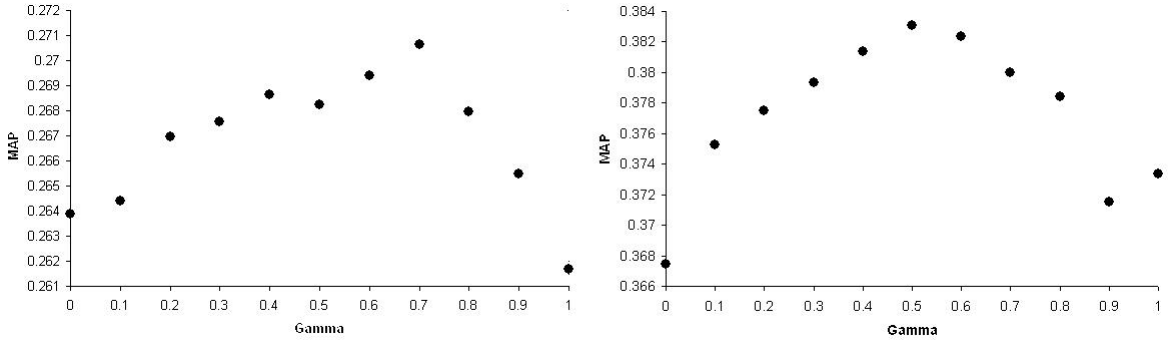

Figure 3: Effect of γ on MAP for the TQE approach on the AP (left) and WSJ (right) test data sets.

topic data is ordered from left to right in ascending MAP score achieved by the RM3 approach.

These graphs suggest that the TQE approach does not outperform RM3 on every test topic. The topics which TQE boosted the average precision score the most over RM3 were TREC topic 159: *Electric Car Development* (on AP) and TREC topic 156: *Efforts to enact Gun Control Legislation* (on WSJ). The TQE approach hurt the AP score the most for TREC topics 172: *The Effectiveness of Medical Products and Related Programs Utilized in the Cessation of Smoking* (on AP), and 157: *Causes and treatments of multiple sclerosis (MS)* (on WSJ).

It was noted that even though TREC topic 157 was hurt the most on the WSJ data set, this same topic had the second largest percent increase in average precision over RM3 on the AP data set. This indicates the importance of the TQE vocabulary in determining the expansion terms. An analogy can be made with the results you may receive if you asked two people, one who had read only the Associated Press news articles and the other solely the Wall Street Journal articles, what type of articles they believed a query like: *Causes and treatments of multiple sclerosis (MS)* should return. The difference in their responses may be attributed to the fact that the WSJ corpus is not as likely to contain articles on multiple sclerosis, and hence may not produce as strong a group of semantic associations for the query terms, especially the more informative terms, such as sclerosis.

This example also highlights the fact that the TQE approach does not account for the information value of specific query terms. Measures, such as *inverse document frequency* (idf), have been shown to improve the performance of information retrieval measures. Testing whether measures like idf can boost retrieval effectiveness of the TQE approach is left for future work.

6.3 Parameter sensitivity

The retrieval effectiveness for various gamma values in equation (17) are shown in figure 3. This graph illustrates that the optimal performance is achieved when both associations are used to create the estimates. This supports our assumption that the task of query expansion is benefited by both syntagmatic and paradigmatic associations. It also leads us to conclude that the TE model of word meaning can be used to model cognitive aspects of the user's real information need that assists the information retrieval process.

7 Conclusions and Future Work

The focus of this paper has been to present the TQE approach, a query expansion technique set in the relevance

modelling framework that is underpinned by a cognitively motivated model of word meaning. The TQE approach formally builds an efficient semantic space that augments the query model using vocabulary based semantic features. The TQE approach was able to significantly outperform a unigram relevance model on two TREC newswire data sets.

The findings of this research also suggest that word order and co-occurrence information stored in the TE model can effectively model the types of word meanings that may underpin the user's real information need. We believe our vocabulary based approach, along with strong linguistic and cognitive motivation, adds weight to the growing number of successful query expansion approaches using term dependencies [10, 9].

An evaluation of TQE on larger document collections and in comparison to some of the other dependency based approaches is the next step in comparing and contrasting the key features of the TQE approach.

An area for future work includes extending the binding operation within the TE model, that underpins the TQE approach, to capture information relating to 3-tuples. Recent research on creating distributed memory models using higher order tensors has shown useful property associations may be captured when 3-tuples and 3-grams are considered [2].

References

[1] Jing Bai, Dawei Song, Peter Bruza, Jian-Yun Nie and Guihong Cao. Query expansion using term relationships in language models for information retrieval. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 688–695, New York, NY, USA, 2005. ACM.

[2] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, Volume 36, pages 673–721, 2010.

[3] P. D. Bruza and D. Song. Inferring query models by computing information flow. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 260–269, New York, NY, USA, 2002. ACM.

[4] Norman N. Holland. *The Critical I*. Columbia University Press, New York, USA, 1992.

[5] Michael N. Jones and Douglas J. K. Mewhort. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, Volume 114, pages 1–37, 2007.

[6] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, Volume 104, pages 211–240, 1997.

[7] Victor Lavrenko and W. Bruce Croft. Relevance-based language models. *In Proceedings of the 24th Annual ACM Conference of Research and Development in Information Retrieval*, pages 120–127, 2001.

[8] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments and computers*, Volume 28, pages 203–208, 1996.

[9] Yuanhua Lv and ChengXiang Zhai. Positional relevance model for pseudo-relevance feedback. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 579–586, New York, NY, USA, 2010. ACM.

[10] Donald Metzler and W. Bruce Croft. Latent concept expansion using markov random fields. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 311–318, New York, NY, USA, 2007. ACM.

[11] Stefano Mizzaro. How many relevances in information retrieval? *Interacting With Computers*, Volume 10, pages 305–322, 1998.

[12] Charles A. Perfetti. The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, Volume 25, pages 363–377, 1998.

[13] Reinhard Rapp. The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, pages 1–7, Morristown, NJ, USA, 2002. ACL.

[14] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. Prentice-Hall, 1971.

[15] Magnus Sahlgren, Anders Holst and Pentti Kanerva. Permutations as a means to encode order in word space. In V. Sloutsky, B. Love and K. Mcrae (editors), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1300–1305. Cognitive Science Society, Austin, TX, 2008.

[16] Hinrich Schütze. Word space. In *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann, 1993.

[17] Michael Symonds, Peter Bruza, Laurianne Sitbon and Ian Turner. Modelling word meaning using efficient tensor representations. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC25)*. In Press, 2011.

[18] Peter D. Turney and Patrick Pantel. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, Volume 37, pages 141–188, January 2010.

[19] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM.

[20] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, New York, NY, USA, 2001. ACM.